Indago: A Novel Search and Analysis Tool for Electronically Stored or Transmitted Content

Jorge H. Román, HPC-1; David H. DuBois, HPC-5; Shelly Spearing, HPC-1; Andrew J. DuBois, Mike Boorman, HPC-5; Ekaterina A. Davydenko, HPC-1; Carolyn M. Connor, HPC-5; Robert L. Gurule, HPC-1

As data generation and storage technologies have advanced, society itself has become increasingly reliant upon electronically generated and stored data. Digital content is proliferating faster than humans can consume it. Current search/filter technology is well suited for simple searches/matches, but a more powerful paradigm is required for complex searches such as finding sensitive corporate knowledge that may be flowing in the intranet and could be accidentally or maliciously sent out over the internet. Context-based search and analysis can greatly enhance the exploration of data and phenomena by essentially reducing the data deluge and increasing the efficiency and effectiveness of the human analyst and/or end-user to access and fully exploit the data-to-knowledge potential that is inherent but latent in nearly every collection.

As technology has continued to advance, modern society has become increasingly reliant upon electronically generated and stored data and information. Digital archives are growing everywhere both in number and in size. Correspondingly, the need to process, analyze, sort, and manipulate data has also grown tremendously. Researchers have estimated that by the year 2000, digital media accounted for just 25% of all information in the world. After that, the prevalence of digital media began to skyrocket, and in 2002, digital data storage surpassed non-digital for the first time. By 2007, 94% of all information on the planet was in digital form [1]. The task of processing data can be complex, expensive, and time-consuming. Applications that alleviate the processing burden and allow users to access and manipulate data faster and more effectively to cross the data-to-knowledge threshold to enable informed, actionable decision-making, particularly for large data streams or digital repositories, are in demand.

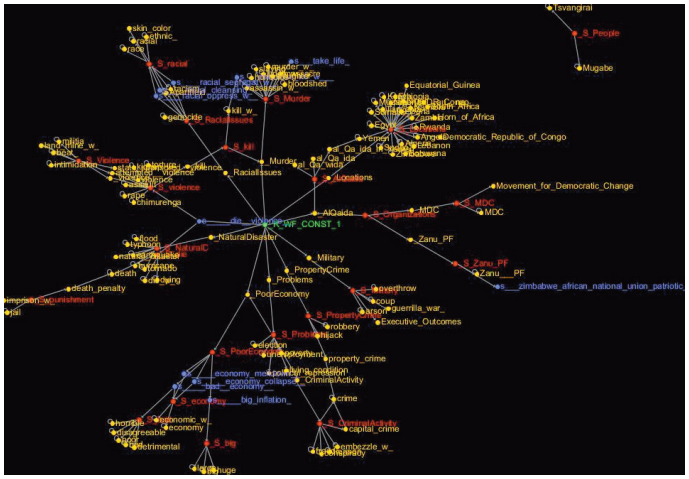


Fig. 1. Graphical illustration of complex, context-based, relationships and connectivity of decision rule set.

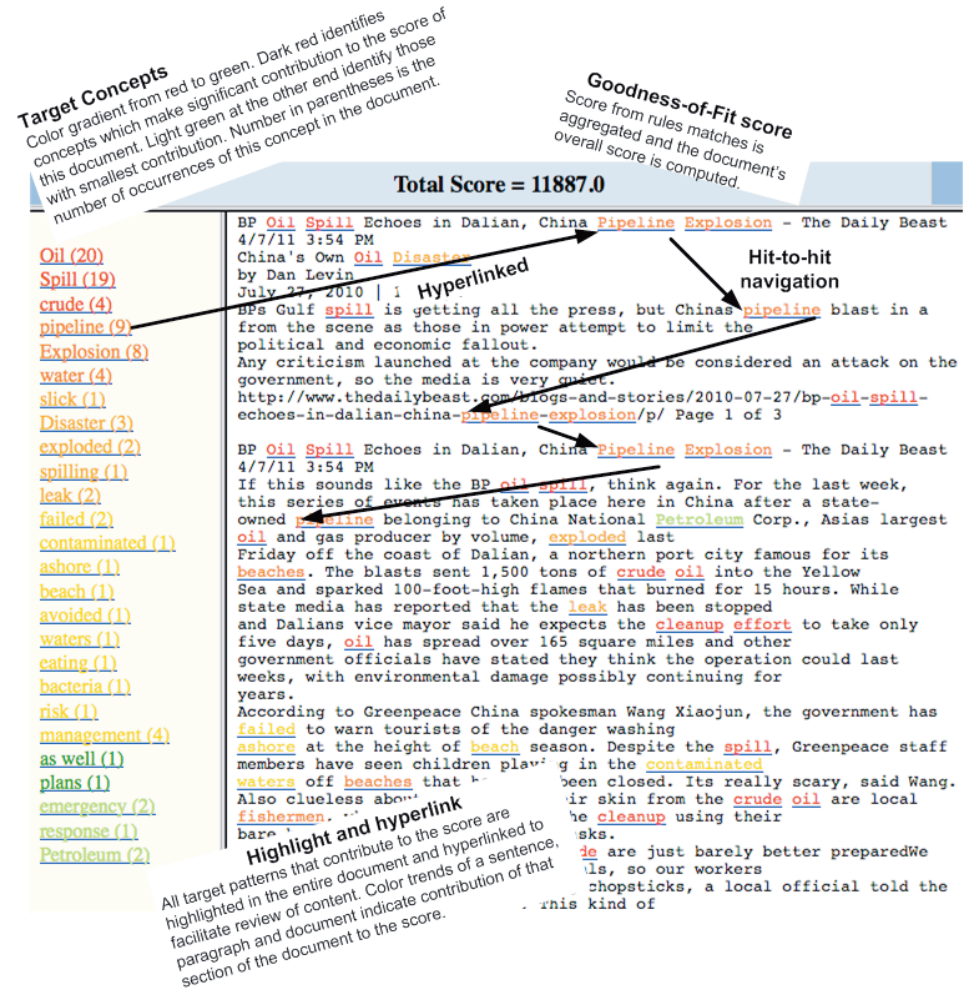
Indago's primary function is the contextual analysis of large repositories of electronically stored or transmitted textual content. A key difference from current approaches is the weighted identification of "concepts in context." Most filters and search engines use either a simple word list or Boolean logic to express desired patterns. For example, Google.com uses a simple list, augmented by additional data (e.g., prestige of pages linking to the document, page ranking, etc.), to produce good retrieval rates; however, the number of matches can be impracticably high with many false positives.

Indago's contextual analysis allows creation of complex models of the language contained in unstructured text in order to build increasingly sophisticated tools that move beyond word occurrence, enabling Indago to make connections and discover the patterns found in large collections.

Indago computes the relevance (or goodness-of-fit) of the text contained in an electronic document based on a predefined set of rules created by a subject matter expert. These rules are modular and are expressed as hierarchical concepts in context. The rules can be nested and are complex in nature to encode the subject matter expert's interpretation of a target concept, such as sensitive corporate knowledge. The system uses a combination of software and hardware to achieve near-real-time analysis of large volumes of text. The currently deployed implementation is used as a context filter for an email server. However, the technology has broad applicability, as the need for fast and accurate search, analysis, and/or monitoring of digital information transcends industry boundaries.

A second key difference from current approaches is that Indago provides a unique, hardware-assisted solution that exploits commercial off-the-shelf (COTS) hardware end-to-end. A fully operational software-based contextual analysis tool has been successfully deployed in a production

For more information contact Carolyn Connor at connor@lanl.gov



Screenshot of Indago output. Example analysis of a publically available news article (Dan Levin, "China's Own Oil Disaster," The Daily Beast, www.thedailybeast.com, July 27, 2010).

environment. In addition, a preproduction system has been developed to demonstrate hardware-accelerated capability.

The hardware-accelerated design enables the inspection of every bit of data contained within a document at rates up to 20 Gbps, far exceeding the capability of current generation multi-core processors while consuming substantially less energy, making Indago a high-performance, low power, green solution. Indago is an inline service as compared to implementations that require large computer clusters with high-end machines. Indago's current hardware acceleration is based on Netlogic technology. Netlogic technical specifications quote power consumption at ~1 Watt per 1 Gbps of processing speed; therefore, at the full rate of 20 Gbps of contextual processing, estimated power consumption would be 20 Watts, which is better than the power consumption of a single computer node by at least an order of magnitude. Comparison to a cluster of compter nodes, as some competing approaches require, is far more impressive. While this technology is designed for and widely applied in the data communications industry, Indago is the first to apply Netlogic's deep-packet processing technology to the field of digital knowledge discovery.

Indago is a cost-effective solution that provides unparalleled performance and capability using proven COTS. It allows users (e.g., scientists, information technology personnel, law firms, etc.) to engage their data faster, more accurately, and more effectively, thus allowing them to solve problems faster, more creatively, and more productively. Furthermore, Indago is domain-adaptable, power-efficient, and fully scalable in terms of rule set size and complexity.

[1] Martin, H. et al., *Science* **332,** 60 (2011).

Funding Acknowledgments
DOE